UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Development of a new pharmacogenomics system for lung cancer based in next generation sequencing

Lúcia de Mendonça Heitor

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

Versão Pública

Dissertação orientada por:
Prof.ª Doutora Margarida Henriques da Gama Carvalho

2018

To everyone that helped me through this journey.

Acknowledgements

Now that this work is complete, I'd like to thank the many people that supported me all through this one year of work. Without some them this would never have been possible, while for others it would still be done but differently.

First, I'd like to thank my advisor for the opportunity to work on this dissertation: I hope it was worth it. Thank you for all the pushes and for reminding me that I still had work to do. I definitely needed those to finally get to this stage.

I'd also like to thank my family for being worried about me being always on the computer. I'm sorry for procrastinating a lot when I claimed to be working. Thank you for putting up with me and my rants about the never-ending problems I always had to solve during this year.

A special thank you to my cats: Xaneco, Mi, Simba, Spot, Pinta and Pussy. To the first five, thank you for letting me hold and snuggle while complaining about life. To the last one, thank you for singing in agony to leave the house while I was writing this document.

Another thanks goes to the people sharing the office with me. Sorry for the unfunny jokes about me never getting this dissertation done and for always giving me condescending looks whenever I'd postpone its writing. Your help, support and lunch company led me here so I owe you all a lot.

Thank you, my classmates, for always making me feel better about whatever I was going through all these days of the second year of Master's. Despite the odds, we motivated each other to continue working while finding ways to laugh at our misery. All is fine, and will be.

Finally, a thank you to my boyfriend. You're the only one who had to go through the compilation of everything the others had to handle from me. Thank you for all the patience you had to have, but you were already obliged to so you can't complain.

To the jury that will read my work and other people who may stumble upon it: I hope it isn't so bad.

Resumo

A medicina de precisão propõe uma nova abordagem médica que enfatiza o papel das aberrações moleculares como causas para a ocorrência de doenças. Por agora, esta iniciativa está a concentrar o seu investimento na oncologia de precisão. O objectivo é oferecer um diagnóstico preciso a um estado adiantado da doença e possibilitar a identificação de uma terapia adequada usando informação relativa ao genoma, estilo de vida e ambiente em que o doente se insere. A aplicação da medicina de precisão permitirá a prevenção de doenças, a redução dos custos associados, e a toxicidade dos tratamentos. O cancro do pulmão é o que mais beneficia desta nova abordagem. Este tipo de cancro é responsável pelo maior número de mortes globais relacionadas com este tipo de doença, com uma taxa de sobrevivência de cinco anos de 18%. O carcinoma do pulmão de células não pequenas é o tipo de cancro do pulmão mais observado, afetando tanto fumadores como não fumadores. A falta de sintomas nas fases mais adiantadas da doença é a principal razão para o número de mortes causadas, tornando o diagnóstico por perfil molecular a solução mais vantajosa. A terapia direccionada é parte das estratégias de tratamento da medicina de precisão em que os fármacos actuam directamente no gene ou proteína com a mutação responsável pelo crescimento do cancro. Como parte do consórcio LungCARD Rise, um procedimento de análise de dados genómicos foi desenvolvido para detectar alterações patológicas em dados de sequenciação recolhidos de amostras de sangue. O uso deste tipos de amostra para detectar cancro contrasta com a biópsia sólida, o procedimento actualmente usado, em que a primeira é menos invasiva, repetível e pouco dispendiosa. As amostras são introduzidas no chip LungCARD desenvolvido que captura células tumorais circulantes e prepara o material genético destas para sequenciação. Este procedimento de análise consegue detectar polimorfismos de um nucleótido, deleções, inserções e a mistura destas duas últimas alterações a um limite de deteção de 1%, reportando-as num relatório automaticamente escrito com informações adicionais para ajudar o médico a decidir a terapia direcionada mais eficaz para um doente com um carcinoma do pulmão de células não pequenas.

Palavras Chave: Carcinoma de pulmão de células não pequenas, Sequenciação de próxima geração, Medicina de precisão, Consórcio LungCARD Rise

Abstract

Precision medicine proposes a new approach to medical procedures as it emphasizes the role of molecular aberrations as the cause for the occurrence of diseases. For now, the investments for this initiative are being focused on Precision Oncology. The objective is to provide a precise diagnosis at an earlier stage and adequate therapy using the patient's genomic, lifestyle and environmental data. The application of precision medicine may allow the prevention of many diseases while also reducing the costs and toxicity of therapy. Lung cancer is one of the cancers benefiting the most from this new framework. This disease is responsible for most deaths caused by cancerous diseases worldwide, with a five year survival estimate of 18%. Non-small cell lung cancer is the most prominent type of lung cancer which affects both smokers and non-smokers. The lack of symptoms at early stages is the main reason for the high number of deaths, making diagnosis through molecular profiling the best method for detection. Targeted therapy is part of Precision Medicine's treatment strategies where the drugs administered act specifically on the precise mutated gene or protein responsible for the cancer growth. As part of the LungCARD Rise Consortium, an analysis workflow was developed to detect and report pathologic alterations on the sequencing data retrieved from blood samples. The use of blood to detect cancer contrasts with solid biopsy, the current procedure, as it's less invasive, repeatable and cheaper. These samples are introduced on the developed LungCARD chip that captures circulating tumor cells and prepares the DNA from these for sequencing. This workflow can detect single nucleotide polymorphisms, deletions, insertions and indels with a limit of detection of 1%, reporting these on a automatically written report with additional information to help clinicians decide the most effective targeted therapy to provide to a patient suffering with non-small cell lung cancer.

Keywords: Non-small cell lung cancer, Next Generation Sequencing, Precision medicine, LungCARD Rise consortium

Resumo Alargado

A medicina de precisão é a nova aposta para a melhoria dos serviços médicos. Esta abordagem pretende utilizar dados genómicos, ambientais e relativos ao estilo de vida para mais facilmente prever se um indivíduo se encontra predisposto a contrair uma determinada doença e tentar combatê-la o mais rapidamente possível. Um dos principais pilares da medicina de precisão é a associação de anomalias genéticas com o desenvolvimento de doenças, recorrendo à elaboração de perfis moleculares como forma de diagnóstico. Para obter estes perfis, o material genético do indivíduo é sequenciado. É neste ponto que a medicina de precisão é associada às tecnologias de Sequenciação de Próxima Geração (NGS), que permitem obter informação sobre o código genético dos indivíduos com elevada precisão. Um ramo da medicina de precisão é a farmacogenómica, que se define como o estudo da eficácia de determinados fármacos segundo o perfil molecular da pessoa. Como consequência destes estudos surgem as terapias direcionadas, que atuam diretamente no(s) gene(s) ou na(s) proteína(s) causadora(s) de doença, impedindo dessa forma o progresso da patologia.

O cancro do pulmão é o cancro que mais mortes causa a nível global e apresenta uma taxa de sobrevivência de cinco anos de 18%. Estes valores devem-se à ausência de sintomas durante as fases iniciais da doença, surgindo apenas quando o cancro já se encontra em fase avançada. Por esta razão a medicina de precisão surge como uma forma de diagnosticar precocemente este cancro. A deteção de alterações características do início do crescimento do tumor permite uma atuação rápida e eficaz na sua prevenção. Por outro lado, a aplicação de terapias direcionadas oferece uma alternativa à quimioterapia com efeitos secundários menos debilitantes. Estas terapias têm como objetivo proporcionar um tempo de sobrevivência prolongado com a melhor qualidade de vida possível para doentes com cancro do pulmão em estado avançado. Este cancro apresenta duas histologias diferentes que o separa em dois tipos: carcinoma do pulmão de células não pequenas e o de células pequenas. O primeiro é o mais proeminente de todos os casos de cancro do pulmão e pode aparecer tanto em fumadores como não fumadores.

Nesta dissertação foi desenvolvido um procedimento de análise de dados genómicos para a deteção de aberrações genéticas que podem estar relacionadas com o desenvolvimento e resistência ao tratamento do carcinoma do pulmão de células não pequenas. Este projeto está inserido no contexto do consórcio LungCARD Rise onde foi desenvolvido um chip que captura células tumorais presentes no sangue. O procedimento utilizado hoje em dia para a obtenção de ADN

tumoral para diagnóstico recorre à obtenção de um pedaço do tumor, normalmente retirado por cirurgia. Este é designado por biópsia sólida e apresenta várias desvantagens na aplicação da medicina de precisão. A deteção de conteúdo genético tumoral a partir do sangue (biopsia líquida) permite que sejam recolhidas amostras ao longo do tratamento de uma forma menos invasiva e dispendiosa. O recurso a cirurgia para obter ADN tumoral não só não pode ser aplicado a todos os doentes, como normalmente é executado apenas uma vez e o que é recolhido é primeiro submetido a análises de anatomia patológica. Estas análises envolvem a utilização de formalina e parafinização da amostra o que degrada o ADN contido nesta. A introdução de amostras de sangue no chip LungCARD permite a captura de células tumorais libertadas pelo tumor (circulantes), e o seu conteúdo genético é extraído. As regiões de interesse do ADN tumoral são amplificadas por PCR e preparadas para sequenciação. Os dados obtidos serão então analisados através da metodologia definida neste projeto, de forma a que no fim as variantes detetadas nas amostras e que se encontrem associadas a este tipo de cancro sejam apresentadas num relatório final. Este terá por sua vez o objetivo de oferecer os resultados da análise dos dados de um doente específico, de forma a facilitar a decisão relativa à terapia direcionada a aplicar.

Para o desenvolvimento do processo de análise de informação foram usados dados especificamente criados como modelo do que será esperado do chip LungCARD para validar todo o procedimento. Este é composto por vários passos que foram definidos segundo o que é mais adequado para a deteção de variantes no fim do processo. O primeiro passo é o pré-processamento dos dados em que estes são tratados e filtrados de forma a reter apenas as sequências de elevada qualidade. A seguir, estas são alinhadas à referência usada de forma a saber a sua posição genómica e para mais tarde permitir a deteção de variantes. Os alinhamentos são igualmente processados, eliminando os de má qualidade, os marcados como alinhamento secundário ou suplementar e ainda as sequências não alinhadas. De seguida, as variantes são detetadas por comparação dos alinhamentos com a referência. Se existirem diferenças entre as duas de forma frequente, a variante detetada é considerada uma variante verdadeira. Nem todas as variantes recolhidas são verdadeiras mutações presentes na amostra e estas devem ser filtradas. Para este fim, e ainda para a anotação e escrita do relatório final, foi preparado um *script* em Python, que necessita de um ficheiro contendo todas as variantes conhecidas do cancro em causa. As variantes detetadas também contidas neste ficheiro serão adicionadas ao relatório final, enquanto que as restantes serão incluídas noutro ficheiro em separado. Desta forma evita-se que falsos positivos e variantes desconhecidas dificultem a tomada de decisão do médico responsável. Este fator é importante pois a capacidade de deteção de variantes de baixa frequência é necessária devido à natureza do material biológico recolhido. O ficheiro em que estas são escritas será anotado com ferramentas especializadas, de forma a perceber se uma variante importante não

foi incluída no relatório ou para mais tarde identificar variantes desconhecidas com relevância no desenvolvimento e tratamento do cancro. Com o uso do ficheiro de variantes conhecidas, será possível executar uma filtragem de variantes, dando especial destaque àquelas com informações relevantes para o tratamento a efetuar. Variantes que devem ser incluídas no relatório deverão ser introduzidas neste ficheiro de forma a serem incluídas em futuros relatórios quando detetada.

Existem algumas características que diferenciam este processo de outros aplicados para o mesmo fim. A principal diferença é a utilização de uma referência diferente da referência humana completa. Um estudo anterior concluiu que esta estratégia introduzia erros na análise, mas este efeito não foi verificado quando testado para este fim. Como as regiões amplificadas por PCR são conhecidas, foi possível reduzir a referência humana a estas, tornando o processo de alinhamento das sequências mais rápido. Outra opção tomada foi a utilização da filtragem por qualidade das sequências em vez do aparo destas por qualidade. Foi verificado que com a utilização deste último processo levava ao encurtamento das sequências alvo devido a regiões homopoliméricas. Como é aplicada uma filtragem por tamanho para remover sequências demasiado pequenas, este aparo causa a perda de sequências alvo durante este passo. Com a filtragem por qualidade média mantêm-se estas sequências de boa qualidade, sem o comprimento destas ser reduzido indevidamente, o que poderia levar à perda de variantes verdadeiras. Finalmente, não foi aplicado o processo de recalibragem de valores de qualidade de bases que melhora a precisão da deteção de variantes. Os benefícios da utilização deste passo são controversos, pelo que este passo não foi aplicado neste projeto, mas será uma possibilidade a explorar em investigações futuras.

O cálculo de parâmetros importantes, como a sensibilidade e especificidade do método, foi impossível de realizar devido a problemas nos dados cedidos. Apesar disso, este procedimento é capaz de detetar polimorfismos de um nucleótido, deleções, inserções e a ocorrência destes dois últimos em simultâneo até frequências de 1%. É ainda um método reprodutível e está disponível numa instância da plataforma Galaxy preparada para permitir a análise de dados obtidos pelo chip LungCARD.